EVA2.0: Investigating Open-Domain Chinese Dialogue Systems with Large-Scale Pre-Training

Yuxian Gu*, Jiaxin Wen*, Hao Sun*, Yi Song, Pei Ke, Chujie Zheng, Zheng Zhang, Jianzhu Yao, Xiaoyan Zhu, Minlie Huang, Jie Tang

The CoAI Group, Tsinghua University & BAAI

❖ Technical Framework

Key Techniques	Models	Skills
Data	EVA2.0 Base / EVA2.0 Large / EVA2.0 xLarge	Chit-chat
Model Architectures	#Parameters: 300M / 700M / 2.8B	Model Safety
Pre-Training Approaches	#Layers: 12-12 / 24-24 / 24-24	Commonsense QA
Decoding Strategies	State Dimension: 768 / 1,024 / 2,048	Empathetic, Supportive

- **Key techniques towards a human-like Chinese chatbot**
  - **Data:** **Quality** v.s. Scale
  - **Layer Number:** **Balanced** v.s. Unbalanced Layers
  - **Role Information:** **w/ Role** v.s. w/o Role
  - **Pre-training:** **Train from Scratch** v.s. Further pre-training
  - **Decoding:**
    - Base Method: **Beam Search** v.s. Sampling
    - Length Control: **Length Penalty** v.s. Min Length
    - Handling Repetition: **w/ Control** v.s. w/o Control

❖ Data

- **Data Collection**
  - all data is publicly available

Weibo Comment	**Weibo Repost**	**Zhihu**
Baidu Zhidao	**Subtitle**	**Story Dialog**
Baidu Tieba	**Douban**	**Public Dataset**

- **Data Quality Control**

Coarse Filtering	Fine Filtering
context-level filtering	relevance score (BERT)
sensitive word filtering	fluency score (LM)
...	...

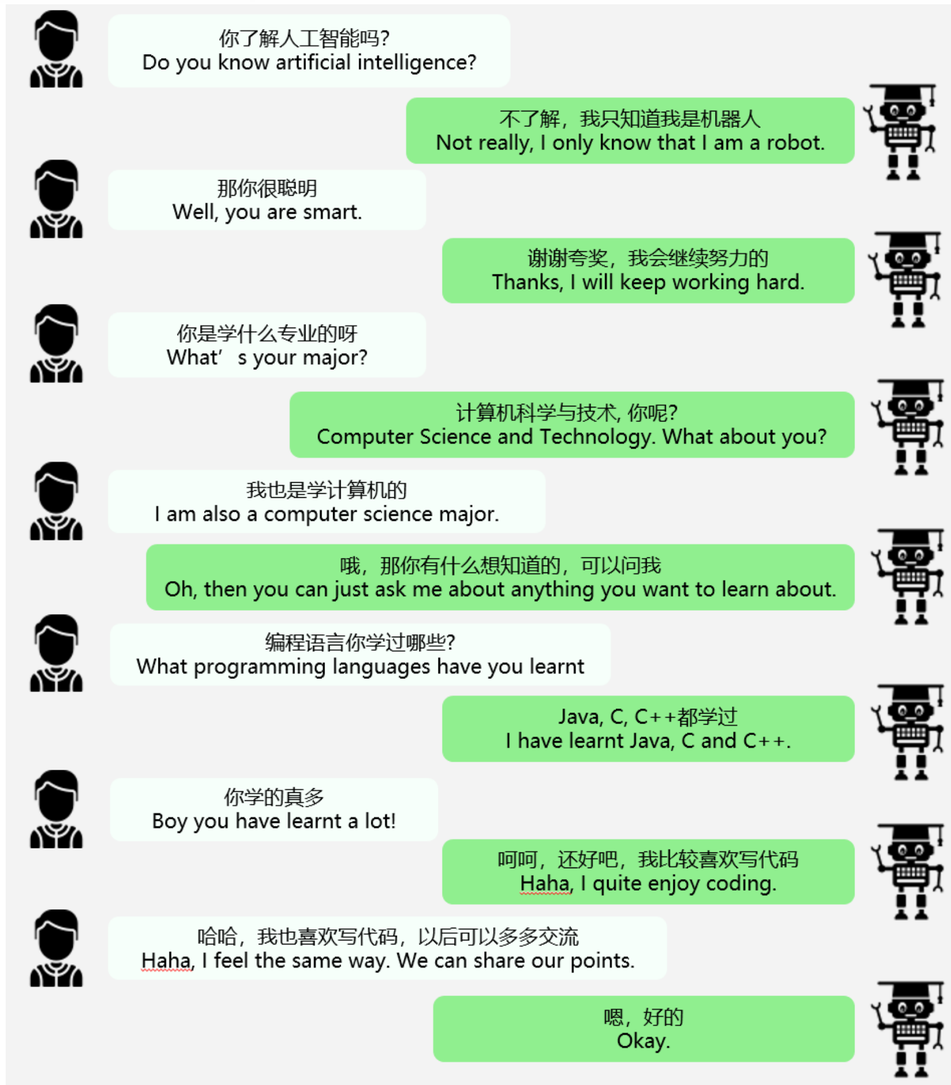
- **Data Analysis**
  - 60GB high-quality dialogue pre-training dataset
  - Basic statistics

Dataset	#Sess.	#Uttr.	#Token	Storage
WDC-Dialogue (Zhou et al., 2021)	1.4B	3.0B	78.3B	181GB
EVA2.0-dataset	0.4B	1.1B	22.4B	60GB

  - Quality statistics

Dataset	Relevance ↑	Fluency ↑	Entertainment ↓
WDC-Dialogue (Zhou et al., 2021)	55.2	-7,147	7.0%
EVA2.0-dataset	93.8	-3,237	6.2%

❖ Case Study



❖ Evaluation

- **EVA2.0 significantly outperforms other open-source counterparts in both automatic and human evaluations.**
- **Strategies Comparison**

Techniques	Model	Single-Turn F1	Single-Turn R-L	Single-Turn B-4	Single-Turn D-4	Multi-Turn F1	Multi-Turn R-L	Multi-Turn B-4	Multi-Turn D-4
Model	6-18	15.6	13.3	1.48	49.4	16.1	13.7	1.54	46.2
	18-6	15.5	13.4	1.52	50.0	16.2	13.9	1.43	45.6
	12-12 ⋆	**16.2**	**13.8**	**1.63**	**53.4**	**16.6**	**14.3**	**1.74**	**50.2**
	+role	13.3	11.3	1.29	45.6	14.4	12.0	1.31	42.3
Pre-training	scratch ⋆	**17.0**	**14.9**	**2.23**	67.7	**17.8**	**15.4**	**2.89**	**66.4**
	further	16.1	13.9	1.77	**68.2**	16.6	14.3	1.84	59.7
Decoding	greedy	16.4	14.1	2.09	63.1	16.5	14.2	2.76	64.2
	sampling	12.2	10.4	1.20	**91.6**	12.5	10.7	1.99	**91.5**
	beam search	16.5	14.7	2.80	43.3	16.9	15.0	3.50	46.0
	+sampling	16.3	14.5	2.21	75.4	16.4	14.6	2.59	73.2
	+len_penalty	**17.4**	**15.4**	**3.23**	66.2	**17.8**	**15.7**	**3.79**	64.9
	+no-repeat ⋆	17.0	14.9	2.23	67.7	**17.8**	15.4	2.90	66.9
	+min_len	16.4	14.2	2.04	62.3	17.1	14.9	2.47	62.8

- **Automatic Evaluation**

Model	Single-Turn F1	Single-Turn R-L	Single-Turn B-4	Single-Turn D-4	Multi-Turn F1	Multi-Turn R-L	Multi-Turn B-4	Multi-Turn D-4
CDial	9.9	8.6	0.67	61.2	11.9	10.3	0.88	63.9
EVA1.0	13.1	11.3	1.27	50.7	15.3	13.2	1.94	56.3
$EVA2.0_{Base}$	16.2	13.8	1.63	53.4	16.6	14.3	1.70	50.2
$EVA2.0_{Large}$	16.4	14.0	1.67	55.8	17.0	14.9	2.03	53.9
$EVA2.0_{xLarge}$	**17.0**	**14.9**	**2.23**	**67.7**	**17.8**	**15.4**	**2.90**	**66.9**

- **Human Evaluation**